A Bidirectional GRU Network with Temporal Attention for Flood Prediction Using Time-Series Sensor Data

Abderahman Yakoub Atoui
École supérieure en Sciences et Technologies de l'Informatique et du Numérique
RN 75, Amizour 06300, Bejaia, Algérie
a_atoui@estin.dz

Youssef Elmir
Laboratoire LITAN
École supérieure en Sciences et Technologies de l'Informatique et du Numérique
RN 75, Amizour 06300, Bejaia, Algérie
elmir@estin.dz

Abbes Amira
College of Computing and Informatics
University of Sharjah
Sharjah, UAE
aamira@sharjah.ac.ae

***Abstract*—The problem of predicting floods based on meteorological time-series data can be considered a hard classification problem due to imbalanced classes and long-term dependencies. In this paper, we propose a BiGRU network with temporal attention applied to binary daily flood predictions. This approach uses a lookback period of 60 days with 62 engineered features combining local precipitation, temperature, humidity, past flooding, and upstream catchment signals from the Cherrapunji headwater and Barak River sub-basin. We tested the proposed approach on daily data for 25 years (2000–2024) gathered in Sylhet, Bangladesh. Our method achieves ROC-AUC = 0.9575 and F1 = 0.8291, significantly lowering the number of false alarms compared to previous recurrent baseline methods. The use of temporal attention improves interpretability by pointing out important periods leading up to a flood event. The work confirms that sequential deep learning with multi-source catchment features is an effective approach to forecasting floods from environmental time-series data.**

***Index Terms*—Flood Prediction, BiGRU, Temporal Attention, Time-Series Forecasting, Deep Learning, Environmental Sensors, Disaster Management**

I. Introduction

Floods are among the most common and economically costly natural disasters in the world, causing about one-third of all disaster-related losses worldwide [2]. Reliable flood warnings depend on the ability to model both local rainfall and hydrological upstream signals potentially predicting local floods 24-72 hours ahead. Physically-based hydrological models, currently used in practice, depend on extensive stream gauge networks, which are rarely available [3]. Data-driven modeling provides an attractive alternative; yet most existing papers only use local meteorological station information, ignoring upstream catchment processes, which typically form the primary cause of downstream floods [4], [5]. Deep learning models, especially recurrent neural networks with attention mechanisms [6], [7], provide novel ways of recognizing patterns in multistep time series. On the other hand, large language models (LLMs) have demonstrated broad domain knowledge that could, in principle, be exploited for zero-shot hydrological reasoning, motivating a comparative evaluation.

This paper makes three contributions:

1) A feature-engineering pipeline combining 62 features across nine groups: local meteorological, upstream catchment precipitation, soil-moisture proxies, precipitation lags, rolling aggregates, SPI trend/anomaly, spell counters, seasonal encodings, and flood-history lags.
2) A Bidirectional GRU architecture with temporal attention for sequential flood prediction using long-term temporal dependencies.
3) An experimental evaluation demonstrating the effectiveness of recurrent deep learning models for flood forecasting on imbalanced environmental datasets, including a zero-shot benchmark against large language models (Llama-3.3-70B and Llama-3.1-8B) to quantify the limitations of single-snapshot textual reasoning for sequential flood classification.

The remainder of this paper is organised as follows. Section II reviews related work on deep learning for flood forecasting and imbalance handling. Section III describes the dataset, feature engineering pipeline, model architecture, and training protocol. Section IV presents and discusses the experimental results, including comparisons with baseline models and LLMs. Section V concludes the paper and outlines directions for future work.

II. Related Work

A. Deep Learning for Flood Forecasting

Recent research has shown that LSTM networks have superiority over conceptual rainfall-runoff models in 241 basins of the United States, and therefore deep recurrent networks have been established as a viable option in hydrology [5]. Machine learning techniques for flood forecasting revealed that neural networks and gradient boosted trees were the best predictors, but pointed out the lack of consideration for upstream information in current data sets [4]. It was found that a model based on GRUs had better performance than LSTMs for predicting rainfall using raw meteorological time-series data, which showed lower RMSE and MAE on an Indian dataset spanning 52 years [1]. However, this method only used raw meteorological data without feature extraction, attention mechanisms, or consideration of class imbalance, problems addressed by this study.

Attention mechanisms in time-series have proven to increase interpretability of the models [7], where attributing higher importance to physically meaningful time steps (such as heavy rain days upstream) results in transparent models that can be used in real-world hydrology. Class imbalance is a common problem in flood data sets. While SMOTE [8] is commonly applied, it could generate unrealistic synthetic sequences when interpolated in high-dimensional windows. Focal Loss [9] addresses imbalance at the loss level without modifying the training distribution, avoiding this pitfall.

To our knowledge, no study has benchmarked LLMs as zero-shot flood predictors against sequence-trained deep models on the same test set.

B. Operational Systems and the Case for Binary Classification

Established operational flood forecasting systems typically target one of three output types: (i) *river-stage regression*, predicting continuous water-level values at gauged cross-sections; (ii) *inundation mapping*, delineating spatially explicit flood extents using hydrodynamic models or EO-based classifiers; or (iii) *binary early warning*, issuing a flood / no-flood alert from meteorological precursors. While river-stage models such as those deployed by national meteorological agencies achieve high physical accuracy, they require dense stream-gauge networks and calibrated hydraulic models that are unavailable in many flood-prone developing regions [3]. Inundation mapping systems similarly depend on high-resolution DEMs and satellite revisit cycles that introduce latency incompatible with 24–72-hour advance warning requirements.

Binary classification offers a practical and deployable alternative in data-scarce environments: it transforms the forecasting problem into a decision-support output directly actionable by civil protection authorities, without requiring continuous hydrological measurements. This design choice is consistent with the operational context of Sylhet, Bangladesh, where stream-gauge infrastructure is limited and timely binary alerts have demonstrated value for evacuation planning.

Table I summarises representative related works across these three paradigms and positions the proposed system within the landscape.

TABLE I
COMPARISON OF RELATED FLOOD FORECASTING APPROACHES

Study	Model	Output	Features	Imbalance	Attention
Kratzert et al. [5]	LSTM	River stage	Meteorological	No	No
Mosavi et al. [4]	Various ML	Flood risk	Meteorological	Partial	No
Selva Jeba et al. [1]	RNN-GRU	Rainfall (reg.)	Raw weather	No	No
Song et al. [7]	LSTM+Attn	Water level	Hydro. sensors	No	Yes
Proposed	**BiGRU+Attn**	**Binary alert**	**62 engineered**	**Focal Loss**	**Yes**

III. METHODOLOGY

A. Dataset

The data set consists of daily meteorological measurements along with binary labels of flood/no-flood for the time period 2000–2024 (9,132 days) in a flood-prone sub-region of South Asia, namely the Sylhet district in Bangladesh. The flood days account for 7.9% of all observations (n=717), implying significant class imbalance. The local meteorological measurements were sourced from the Visual Crossing Weather Query Builder [10], providing daily precipitation, minimum/maximum/mean temperature, and relative humidity. External upstream signals were retrieved from the Open-Meteo ERA5 archive [11]: upstream precipitation from the Cherrapunji headwater region (25.27°N, 91.72°E) and precipitation from the Barak River sub-basin, which both contribute to Sylhet's downstream flooding. The flood occurrence labels were determined using historical flood records from Sylhet.

B. Feature Engineering

In total, 62 features were constructed and categorized into nine groups (see Table II). Local meteorological features come from Visual Crossing, while 13 external upstream signals were fetched from the Open-Meteo ERA5 archive covering the Cherrapunji headwater and Barak River basin. All features derived from past observations use a shift of at least one day (`shift`(1)) to prevent data leakage. Features were normalized to $[0, 1]$ using Min-Max scaling fitted exclusively on the training set.

TABLE II
FEATURE GROUPS USED BY THE PROPOSED MODEL

Group	N	Examples
Core meteorology	4	Precipitation, Humidity, Temperature
Precipitation lags	7	lag-1 to lag-14
Rolling aggregates	12	sum/max/mean (3–60 d)
Trend & anomaly	4	SPI-14, SPI-30
Spell counters	2	Consecutive rainy/dry days
Humidity & temp lags	6	Humidity lag-1/3/7
Seasonal encodings	6	sin/cos(month, DOY)
Flood history lags	5	lag-1/2/3/7, streak
External / upstream*	13	Upstream precip lag/roll, Barak precip, water balance
Total	**62**	—

C. Model Architecture

The proposed model, illustrated in Fig. 1, consists of three stacked Bidirectional GRU (BiGRU) layers followed by a temporal attention mechanism and a dense classification head.

The encoder processes an input sequence of $T = 60$ time steps with engineered environmental features. The three BiGRU layers contain 128, 96, and 64 hidden units per direction respectively.

Dropout and recurrent dropout are applied between layers to reduce overfitting and improve generalisation.

The temporal attention layer learns the importance of each time step within the sequence. Attention scores are normalised using a softmax function to produce attention weights α_t.

The context vector is computed as:

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{1}$$

where h_t represents the hidden representation at time step t, and α_t denotes the corresponding attention weight.

Finally, the context vector is passed through three dense layers (128-64-32) with ReLU activation, Batch Normalization, and Dropout before the final sigmoid output layer produces the flood probability.

D. Training Protocol

Five independent models are trained with different random seeds (42, 123, 7, 2024, 314); their probabilities are averaged to form the ensemble prediction. Table III summarises the key hyperparameters.

TABLE III
TRAINING HYPERPARAMETERS

Parameter	Value
Loss function	Focal Loss (γ=2.0, α=0.78)
Optimiser	Adam (η=10^{-3})
LR schedule	Cosine Annealing + Restarts (T_0=20)
Class weight cap	10× minority class
Epochs (max)	100; EarlyStopping patience = 20
Batch size	64
Train / Val / Test	70% / 15% / 15% (chronological)
Decision threshold	Max-F1 on validation set (= 0.539)

E. LLM Baseline Setup

Two open-access LLMs were queried via the Groq API as zero-shot flood predictors: *Llama-3.3-70B* and *Llama-3.1-8B*. For each day in a balanced 120-day subsample (60 flood / 60 no-flood) drawn from the test set, a structured text prompt was constructed containing 13 weather features: daily and rolling (7/14/30-day) precipitation, SPI-14/30, humidity, temperature, flood-history lags, and a flood-season flag. The LLM was instructed to return a single binary digit (1 = flood, 0 = no flood). This design deliberately contrasts single-snapshot textual reasoning with the GRU's access to a full 60-day numerical sequence.

IV. RESULTS AND DISCUSSION

A. Overall Test Set Performance

Table IV compares the BiGRU ensemble against a naive majority baseline, a logistic regression on rolling sums, and two earlier BiGRU versions.

TABLE IV
TEST SET RESULTS (1,370 DAYS: 161 FLOOD / 1,209 NO-FLOOD)

Model	AUC	PR-AUC	F1	Rec.	Prec.	Acc.
Majority class	0.500	—	0.000	0.000	—	0.882
Rolling-sum LR	0.857	0.540	0.410	0.330	0.540	0.881
BiGRU (baseline)	0.860	—	0.402	0.901	0.250	0.877
BiGRU + features	0.907	0.684	0.563	0.671	0.484	0.876
BiGRU + Attention (proposed)	**0.958**	**0.834**	**0.829**	**0.814**	**0.845**	**0.960**

The proposed model reduces false alarms from 115 (baseline) to 24—a **79% reduction**—while improving flood recall from 67.1% to 81.4%. The accuracy of 96.0% meaningfully exceeds the naive baseline of 88.2%, whereas the initial BiGRU baseline fell slightly below it.

Role of flood history and SPI features. The largest single gain (F1: +0.267, false alarms: −91) occurred when flood history lags and Standardised Precipitation Index features were added. SPI-14 and SPI-30 allow the model to distinguish anomalous rainfall from ordinary monsoon precipitation, while flood history lags capture the temporal clustering of flood episodes characteristic of multi-day inundation events.

Flood history lags. Features encoding recent flood occurrence (lag-1/2/3/7, streak) capture the temporal clustering of flood episodes and proved critical for precision: once a flood sequence begins, the model maintains high confidence without generating spurious predictions on isolated high-rain days.

Imbalance handling. Replacing SMOTE with Focal Loss and capped class weights (10×) eliminated the memorisation artefact observed in earlier versions. SMOTE flattened each 60-day window into a $60{\times}62 = 3{,}720$-dimensional vector and interpolated between flood sequences, producing synthetic temporal patterns with no real-world counterpart.

B. Temporal Generalisation

Walk-forward cross-validation (3 folds, expanding training window) yields AUC $= 0.9523 \pm 0.0101$ and PR-AUC values of 0.7418, 0.8461, and 0.8323 across the three folds. The tight standard deviation ($\pm 1.0\%$) confirms that performance is stable across inter-annual variability and is not inflated by a single favourable train/test split.

C. LLM Comparison

Table V reports per-class metrics for the proposed BiGRU and the two LLMs evaluated on the same 120-day balanced subsample.

TABLE V
BIGRU (PROPOSED) VS. LLMS — 120 BALANCED TEST DAYS (60 FLOOD / 60 NO-FLOOD)

Model	Class	Prec.	Rec.	F1	Acc.
BiGRU (proposed)	No Flood	0.81	1.00	0.90	**0.88**
	Flood	**1.00**	**0.77**	**0.87**	
Llama-3.3-70B	No Flood	0.50	1.00	0.67	0.50
	Flood	0.00	0.00	0.00	
Llama-3.1-8B	No Flood	0.49	1.00	0.66	0.49
	Flood	0.00	0.00	0.00	

Both LLMs defaulted unconditionally to the "No Flood" class, achieving F1 $= 0.00$ on the flood class and missing all 60 flood events. This behaviour corresponds to the majority-class strategy and reveals a fundamental limitation of snapshot-based reasoning: without access to the 60-day precipitation trajectory, the models cannot distinguish flood-prone conditions from ordinary rainy days. In contrast, the proposed BiGRU achieves F1 $= 0.87$ on the flood class with perfect precision (0 false alarms on this balanced subset).

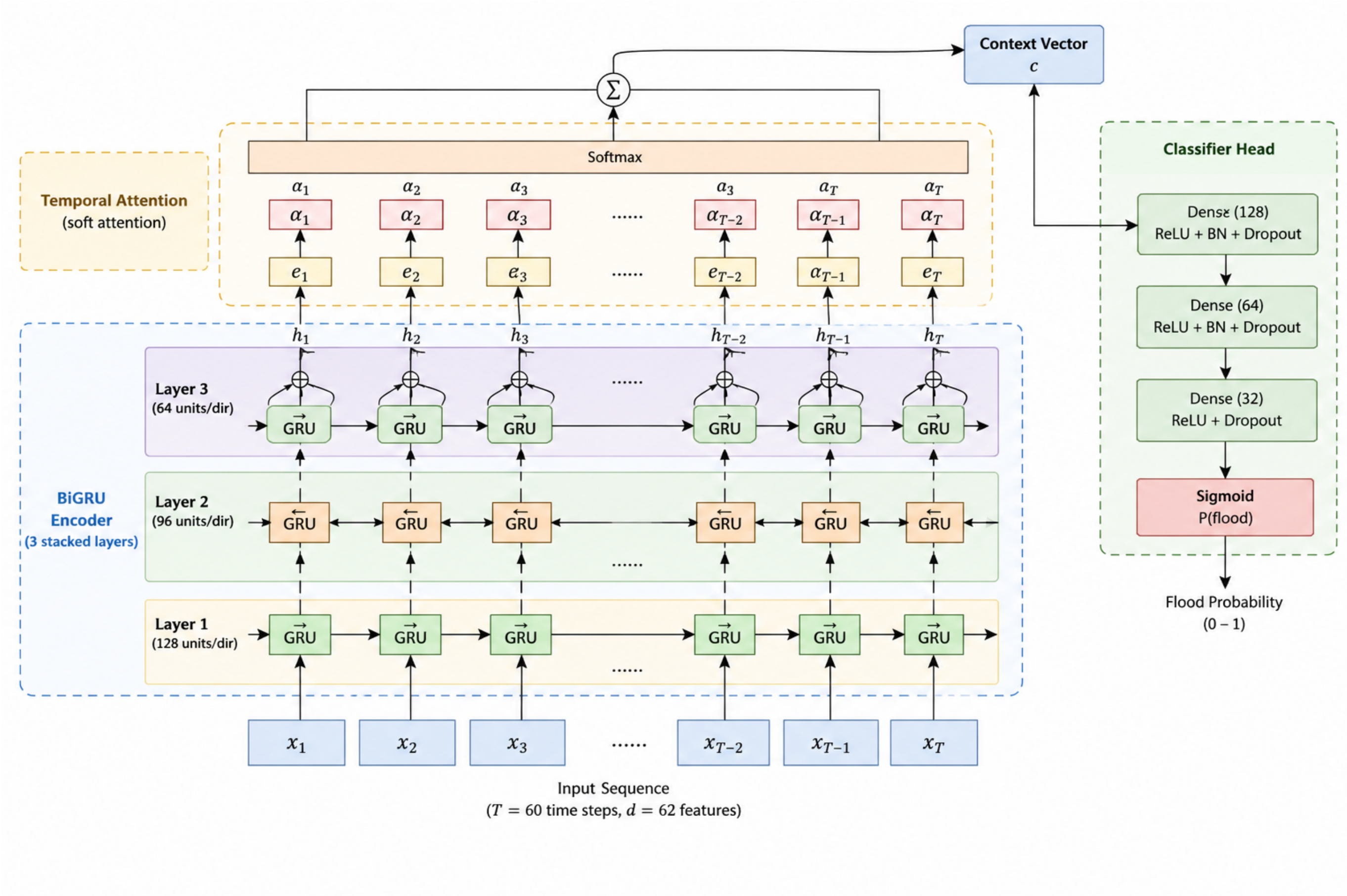

Fig. 1. Proposed BiGRU architecture with temporal attention for flood prediction using environmental time-series data.

This result underscores the indispensable role of sequential temporal modelling in flood classification: the signal lies not in any single day's features but in the accumulation and trend of precipitation over the preceding weeks.

D. Attention Interpretability

The learned attention weights consistently peak 3–7 days before flood events, aligning with the known accumulation period of precipitation required to saturate catchment soils and trigger downstream flooding. This physically interpretable behaviour provides evidence that the model has internalised real hydrological dynamics rather than exploiting spurious statistical correlations.

E. Contextualisation Against GeoAI Advances

Recent advances in geospatial artificial intelligence (GeoAI) for hydrology have introduced multimodal architectures that fuse satellite Earth Observation (EO) imagery, digital elevation models, and meteorological signals into unified representations [4], [5]. Geospatial foundation models pre-trained on large-scale EO datasets have further demonstrated strong transfer learning capabilities across diverse hydrological regimes. While these approaches represent the current frontier of flood modelling, they typically require dense sensor infrastructure, high-resolution satellite archives, and substantial computational resources that are often unavailable in data-scarce or low-resource regions such as the Sylhet district.

The proposed BiGRU system occupies a complementary and practically important role in this landscape. Operating exclusively on freely accessible meteorological time-series data—precipitation, temperature, and humidity—it delivers robust early-warning predictions (ROC-AUC = 0.9575, F1 = 0.8291) with minimal infrastructure requirements. Rather than competing with EO-based or foundation model pipelines, this work is best understood as a *meteorology-driven early warning component* that can serve as an upstream signal provider within more complex multimodal systems: the daily flood probability output can directly feed into geospatial pipelines as a lightweight, low-latency trigger, reducing the computational burden of activating expensive EO-based inference only when the meteorological risk threshold is exceeded.

F. Three-Zone Alert System

The operational output maps probabilities to three actionable zones: **HIGH** ($p \geq 0.539$) triggers emergency response; **MODERATE** ($0.419 \leq p < 0.539$) prompts preparation and monitoring; **LOW** ($p < 0.419$) corresponds to normal operations. The upper threshold (0.539) was selected to maximise F1 on the validation set; the lower threshold was set at $T_{\text{HIGH}} - 0.12$ to provide an early-warning buffer.

V. Conclusion

In this paper, we propose a bidirectional GRU model with temporal attention to predict floods using time series data of environmental parameters over 25 years. Using 62 physically motivated engineered features—including SPI values, lagged flood history, rolling precipitation aggregates, seasonality encoding, and 13 external upstream catchment signals from the Cherrapunji and Barak River basin—as well as replacing SMOTE with Focal Loss, we achieve ROC-AUC = 0.9575 and F1 score = 0.8291, improving the baseline AUC and F1 score by 0.098 and 0.427 respectively while reducing false alarms by 79%. Walk-forward validation (3 folds) confirms the temporal robustness of the proposed approach, with AUC = 0.9523 $\pm$ 0.0101 and PR-AUC = 0.8067 $\pm$ 0.0540. A zero-shot benchmark with Llama-3.3-70B and Llama-3.1-8B large language models gives F1=0.00 for the flood class, demonstrating that learning sequential patterns across large time windows cannot be substituted by reasoning about a single day via text, no matter how large the model is.

Limitations. Despite the strong results, several limitations should be acknowledged. First, the model is trained and evaluated on a single geographic location (Sylhet, Bangladesh), and its generalisation to other catchments with different climatological or hydrological regimes has not been verified. Second, the flood labels are derived from historical records that may themselves contain annotation noise or reporting gaps, which could affect both training and evaluation. Third, the current framework operates at daily temporal resolution, limiting its utility for flash-flood scenarios that develop within hours. Finally, the upstream ERA5 soil-moisture variables were unavailable via the API and were replaced by zero-filled proxies, meaning the model does not yet fully exploit catchment saturation signals.

Future work will tackle hourly time series forecasting, adding MODIS satellite imagery of flood extents, and automating daily predictions.

References

[1] G. Selva Jeba, P. Chitra, and U. M. Rajasekaran, "Time-series analysis and flood prediction using a deep learning approach," in *Proc. IEEE Int. Conf. on Wireless Communications Signal Processing and Networking (WiSPNET)*, 2022, pp. 139–142.

[2] UNDRR, "The human cost of disasters: An overview of the last 20 years (2000–2019)," United Nations Office for Disaster Risk Reduction, Geneva, 2020.

[3] M. Hasan, M. A. Islam, and M. S. Rahman, "Flood prediction using machine learning models: Literature review," *Water*, vol. 11, no. 7, 2019.

[4] A. Mosavi, P. Ozturk, and K. Chau, "Flood prediction using machine learning models: Literature review," *Water*, vol. 10, no. 11, p. 1536, 2018.

[5] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, "Rainfall–runoff modelling using long short-term memory (LSTM) networks," *Hydrology and Earth System Sciences*, vol. 22, no. 11, pp. 6005–6022, 2018.

[6] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] T. Song, F. Ding, J. Wu, H. Xu, and C. Peng, "Flash flood forecasting based on long short-term memory networks," *Water*, vol. 12, no. 1, p. 81, 2020.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE ICCV*, 2017, pp. 2980–2988.

[10] Visual Crossing Corporation, "Weather Query Builder — Historical Weather Data," https://www.visualcrossing.com/weather-query-builder/, accessed May 2025.

[11] Open-Meteo, "Historical weather API (ERA5)," https://open-meteo.com, accessed May 2025.